Simultaneous estimation of multiplicity of infection and allele frequencies from complex malaria infection genetic data

by

Maxwell R Murphy

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Arts

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bryan Greenhouse, Co-chair
Professor Sandrine Dudoit, Co-chair
Professor John Marshall
Professor Rasmus Nielsen

Spring 2019

Simultaneous estimation of multiplicity of infection and allele frequencies from complex malaria infection genetic data

Copyright 2019
by
Maxwell R Murphy

Abstract

Simultaneous estimation of multiplicity of infection and allele frequencies from complex malaria infection genetic data

by

Maxwell R Murphy

Master of Arts in Biostatistics

University of California, Berkeley

Professor Bryan Greenhouse, Co-chair

Professor Sandrine Dudoit, Co-chair

Malaria genetics and genomics offer insight into population structure and epidemiology, and can provide important information that will inform elimination efforts. Estimation of relevant environmental and epidemiological factors such as genetic diversity of complexity of infection help elucidate the impact of interventions and the origin of observed infections. However, high prevalence of polygenomic infections renders estimation of such parameters challenging. Previous methods, such as *THE REAL MCCOIL*, have been developed to address these challenges, allowing for estimation of allele frequencies and complexity of infection from single nucleotide polymorphism data. These methods, however, are incompatible with highly informative multi-allelic genetic data, such as microsatellite genotyping or targeted amplicon sequencing technologies. Therefore, we have developed a Bayesian approach allowing for polygenomic samples to estimate complexity of infection and allele frequencies from highly diverse multi-allelic genetic data. We show that our method is able to accurately recover complexity of infection and population allele frequencies from simulations. We also compare our method to standard approaches in the literature, demonstrating the inherent bias of those methods. We also demonstrate our method on field samples, showing epidemiologically relevant differences in estimates as compared to standard methods. Our method is an important addition to understanding malaria epidemiology, allowing for the full utilization of highly diverse and highly informative genetic loci.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last two decades, global malaria control has been extraordinarily successful, with 46 countries reporting fewer than 10,000 cases, and 26 reporting fewer than 100 indigenous cases in 2017 [13]. The challenges these countries face in implementing elimination programs are unique, as they shift from burden reduction to transmission interruption and preventing its re-establishment, with the most important activities being identification of remaining residual transmission foci and preventing reintroduction from neighboring regions that have not yet entered elimination status themselves [8]. As more countries enter elimination settings, there is an increasing need for more precise tools to characterize local and regional transmission dynamics, with parasite genetics being increasingly used to inform these efforts [3, 11]. However, malaria infections in endemic regions are often composed of multiple genotypes due to the phenomena of multiple-strain infections which arise either due to multiple infectious mosquito biting events, or a single biting event that transmits multiple parasites with distinct genotypes [9].

The presence of multiple genetic strains makes the interpretation of genetic data and estimation of population genetic parameters a challenge due to the convoluted genetic signal that arises due to the limited resolution of the assays available [4]. Because of this, analysis of parasite population genetics has often been restricted to monogenomic infections, greatly reducing the richness of the data available, and potentially introducing bias into results and interpretation if monoclonal infections are not representative of the population as a whole. Other basic parameters such as complexity of infection (COI), that is, the total number of distinct strains in circulation, are also often estimated using methods that are inherently biased. For example, it is common practice to estimate COI by taking the maximum number of observed alleles at a single genotyped locus across multiple genotyped loci. It is apparent that such an estimator will be limited by the diversity of the loci available, and will be biased upward in the case of any false positivity to the assay, and worse yet, bias will increase as more loci are included in the genotyping panel. COI is a biologically meaningful parameter, and its relation with genetic diversity as measured by allele frequencies is often used to inform malaria control strategies and their impact, therefore proper estimation and inference is paramount [1, 6, 7, 12].

To address some of these challenges, *THE REAL McCOIL* was developed to estimate COI and allele frequencies from single nucleotide polymorphism (SNP) genotyping data from mixed infections, while allowing for genotyping error, as parameters in a hierarchical model fit by Markov Chain Monte Carlo (MCMC) [5]. However, the restriction to SNP data is a severe one, rendering assays that target highly diverse and consequently highly informative alleles such as microsatellite genotyping and next generation targeted amplicon deep sequencing incompatible. Motivated by this gap in available estimation techniques and software, we propose a hierarchical model that allows for multi-allelic genetic loci to jointly estimate sample complexities of infection and population allele frequencies of a panel of genetic loci. In chapter 2, we describe the observed data structure, our model specification and priors used, and our approach to fitting the model by MCMC. In chapter 3, we apply our model to a simulation and demonstrate its performance in recovering sample complexity of infection and allele frequencies from diverse genetic loci. In chapter 4, we apply our model to a real world data set, and demonstrate how estimates differ compared to the inherently biased methods currently used. In chapter 6, we conclude the thesis with remarks on software availability and how this model may be further expanded.

Chapter 2

Methods

The following sections provide details regarding the notation used, the observed data, a hierarchical model that describes how the data are generated and associated likelihood specification, and our approach to fitting the model using MCMC.

Notation

Table 2.1 describes the notation used in describing the observed data, our model, and process for fitting the model.

Symbol	Meaning
n	Number of samples
l	Number of genetic loci
a_j	Number of alleles at genetic locus j
i	Index for samples
j	Index for genetic loci
k	Index for alleles
$g_{i,j}$	The binary vector indicating whether or not an allele was observed for sample i at genetic locus j
g^*	A numeric vector indicating the true underlying number of strains that contribute each allele
$G^*_{i,j}$	The set of possible genotypes compatible with the observed data $g_{i,j}$
μ_i	The complexity of infection for sample i
π_j	The vector of population allele frequencies for locus j
ϵ^-	The probability of an allele from a distinct parasite strain to not be detected by the genotyping assay
ϵ^+	The probability of an allele to be falsely detected by the genotyping assay
λ	The population mean complexity of infection
S	The total number of importance samples used

Table 2.1: Model Notation

Observed Data Structure

The observed data are the result of a genotyping assay that detects presence or absence of alleles at a set of genetic loci, but does not resolve the number of distinct organisms contributing the allele, or establish any linkage across genetic loci. Examples of such technologies are microsatellite genotyping by capillary electrophoresis, and targeted amplicon (short diverse DNA segments) deep sequencing by next generation sequencing. Further, the observed data are noisy and subject to errors introduced during the genotyping process. This manifests as either alleles detected that are not truly present, or as alleles that go undetected despite being truly present. We represent the observed data as the collection g, indexed by biological sample i and genetic locus j. The element $g_{i,j}$, then, is a binary vector indicating the presence or absence of alleles, as detected by the genotyping assay.

Model

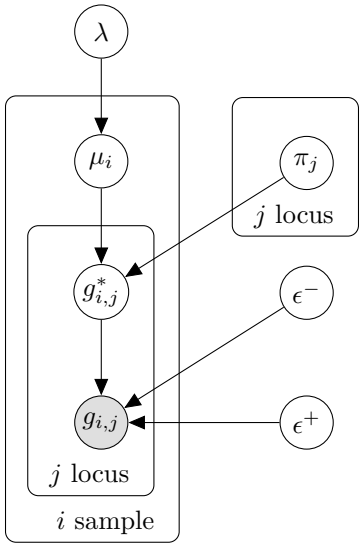

Figure 2.1: Graphical model representation

The boxes are "plates" that represent replicates, filled circles indicate observed data, and empty circles indicate latent variables.

We pose our model as a generative process reflecting the procedure by which infections occur, are collected, and subsequently genotyped. We assume the following generative process:

For each sample i:

1. Choose sample COI $\mu_i \sim$ Zero-Truncated Poisson(λ)

2. For each locus j:

   a) Choose true underlying genotype $g^*_{i,j} \sim \text{Multinomial}(\mu_i, \pi_j)$

   b) Choose $g_{i,j}$ from $p(g_{i,j}|g^*_{i,j}, \epsilon^+, \epsilon^-)$, a vector of independent bernoulli trials for each allele indicating if they are observed or not, conditioned on the true underlying genotype $g^*_{i,j}$ and false positive and negative rates ϵ^+ and ϵ^-

This model is represented as a probabilistic graphical model in Figure 2.1

Posterior Distribution

Given the above model specification, we can express the posterior distribution of sample complexities of infection μ, population mean complexity of infection λ, loci allele frequencies π, and false positive and negative rates ϵ^+ and ϵ^- given the observed data g as:

$$\begin{aligned} p(\mu, \lambda, \pi, \epsilon^-, \epsilon^+|g) &\propto p(\mu|\lambda)p(\lambda)p(\epsilon^-)p(\epsilon^+)\prod_{i=1}^{n}\prod_{j=1}^{l} p(g_{i,j}|\mu_i, \pi_j, \epsilon^-, \epsilon^+) \\ &= p(\mu|\lambda)p(\lambda)p(\epsilon^-)p(\epsilon^+)\prod_{i=1}^{n}\prod_{j=1}^{l}\sum_{g^* \in G^*_{i,j}} p(g_{i,j}|g^*, \epsilon^-, \epsilon^+)p(g^*|\mu_i, \pi_j) \end{aligned}$$

where $G^*_{i,j}$ is the set of all possible true genotypes that could have resulted in the observed data $g_{i,j}$. In the case of an error model where there is positive probability of both false positives and false negatives, as is the case in our model, this is the set of all integer vectors of length a_j (the total number of distinct alleles at locus j) where all elements are greater than or equal to zero, such that the sum of the vector equals μ_i (the complexity of infection in sample i).

We further specify the probability of the observed data $g_{i,j}$ given an underlying genotype g^* as:

$$p(g_{i,j}|g^*, \epsilon^+, \epsilon^-) = \prod_{k=1}^{a_j} \begin{cases} (1-\epsilon^-)^{g^*_k} & \text{if } g_{i,j,k} = 1 \text{ and } g^*_k > 0 \\ (\epsilon^-)^{g^*_k} & \text{if } g_{i,j,k} = 0 \text{ and } g^*_k > 0 \\ (\epsilon^+) & \text{if } g_{i,j,k} = 1 \text{ and } g^*_k = 0 \\ (1-\epsilon^+) & \text{if } g_{i,j,k} = 0 \text{ and } g^*_k = 0 \end{cases}$$

This error model assumes that the probability of a false negative for a particular allele is exponentially less likely the more strains there are that carry that allele. This reflects a belief that the genotyping mechanism is dependent on the relative density of DNA present, and that an allele with more strains contributing is likely to have a higher relative amount and is less likely to "drop out". Conversely, false positives simply occur at rate ϵ^+.

We then specify the probability of an underlying genotype g^* as:

$$p(g^*|\mu_i, \pi_j) = \frac{\mu_i!}{g_1^*! \cdots g_k^*!} \pi_{j,1}^{g_1^*} \cdots \pi_{j,k}^{g_k^*}$$

which is simply the probability of observing g^* as if it were drawn from a multinomial distribution with μ_i trials and event probabilities π_j. This reflects our assumption that observed infections are independent.

Finally, we specify the probability of a particular complexity of infection μ_i as:

$$p(\mu_i|\lambda) = \frac{\lambda^{\mu_i}}{(e^\lambda - 1)\mu_i!}$$

which is the probability of drawing μ_i from a Zero-Truncated Poisson with mean λ. This reflects our belief that all genotyped samples contain at least 1 parasite strain, and that the complexities of infection sampled are drawn from a Poisson distribution with mean λ.

Prior Distributions

For the remaining probability distributions not yet accounted for, we place the following vague prior distributions

$$\begin{aligned} \lambda &\sim Gamma(.25, .25) \\ \pi_1, \ldots, \pi_j &\sim Dirichlet(1_{a.}) \end{aligned}$$

as well as user determined prior distributions for the following

$$\begin{aligned} \epsilon^+ &\sim Beta(\alpha_+, \beta_+) \\ \epsilon^- &\sim Beta(\alpha_-, \beta_-) \end{aligned}$$

Modeling Assumptions

We make several simplifying assumptions in this basic model of parasite transmission and case sampling. First, observed complexities of infection are assumed to be drawn from a Zero-Truncated Poisson distribution. This is not critical and more flexible modeling by other distributions such as a negative binomial is possible. Second, we assume that genetic loci are independent. This is a reasonable assumption as genetic panels can by design have this property. Third, we assume that false positive and false negative rates are independent of genetic loci and sample. This assumption could be weakened in further improvements to the model by accounting for other information available, such as parasite density within sample,

which directly impacts how often alleles go unobserved by genotyping. Fourth, we assume that samples are independent, and that polygenomic infections that occur are from multiple independent infections. This assumption may be violated in environments where genetic diversity is low or transmission is highly localized, causing individuals to be infected by closely related parasites. And finally, we assume that observation of any one allele is independent of observations of other alleles. This assumption is generally acceptable, but could be violated due to limitations of certain genotyping techniques. For example, microsatellite genotyping exhibits a phenomenon known as amplification bias, where shorter fragments of DNA are more readily amplified, potentially drowning out signal from larger fragments. We ignore this as it is highly assay specific, and its impact will be limited to edge cases.

Parameters of Interest

We are interested in estimating π, the population level allele frequencies across loci, μ, the sample specific estimates of complexity of infection, and λ, the population level estimate of mean complexity of infection. From these parameters, we may derive other ecologically relevant statistics such as heterozygosity, a measure of genetic variation, and polyclonal fraction, the percent of samples that have COI > 1.

Inference

To estimate our parameters of interest, namely sample complexities of infection, population mean complexity of infection, and genetic loci allele frequencies, we employ a standard MCMC approach using Metropolis-Hastings and Gibbs sampling steps [10]. We also note that we do not directly estimate the true underlying genotype $g^*_{i,j}$, and instead marginalize over all possible true genotypes. This is done for two reasons. First, estimates of the true underlying counts of alleles are of little interest if they are not used to phase genotypes into their constituent distinct strains. The observed data alone, however, are not sufficient to phase genotypes into distinct strains as there is no information that indicates one set of alleles across genetic loci is more likely than another to be attributed to a distinct parasite strain. The only parameters of interest that would be gleaned from these estimates would be complexity of infection and genetic loci allele frequencies, which we are already directly estimating. Second, even if we were interested in these values, sampling would be very challenging. Because μ_i dictates the support of $g^*_{i,j}$, we would have to simultaneously sample μ_i and g^*_i, which would grow in difficulty as the number of loci in the panel is increased, making scalability of the procedure a challenge.

MCMC Implementation

In this section we give details of our sampling procedure and how we compute the likelihood of the parameters. We have specified prior distributions for our parameters of interest above.

Sampling Procedure

A Metropolis-Hastings algorithm is used to sample from the posterior distribution of μ, π, ϵ^-, and ϵ^+, and λ is sampled by Gibbs sampling. Initialization and sampling is conducted as follows:

COI

We initialize μ to the naive estimate of complexity of infection by calculating the maximum number of observed alleles across all genetic loci for each sample. For each μ_i, we sample changes in μ_i from a geometric distribution with an adaptive parameter that is updated based on whether the previous proposal was accepted or not. The proposed change has its sign reversed with probability .5. The user may provide a maximum COI, in which case values that fall outside of the range $1 \leq \mu_i^* \leq \mu_{max}$ are rejected. Otherwise, we accept the proposal with probability

$$
\begin{aligned}
&= \min\{1, \frac{q(\mu_i|\mu_i^*)p(\mu_i^*|g_{i,\cdot}, \pi, \epsilon^-, \epsilon^+)}{q(\mu_i^*|\mu_i)p(\mu_i|g_{i,\cdot}, \pi_i, \epsilon^-, \epsilon^+)}\} \\
&= \min\{1, \frac{p(\mu_i^*|g_{i,\cdot}, \pi_i, \epsilon^-, \epsilon^+, \lambda)}{p(\mu_i|g_{i,\cdot}, \pi_i, \epsilon^-, \epsilon^+, \lambda)}\}
\end{aligned}
$$

due to symmetry of the proposal distribution.

Allele Frequencies

We initialize π to the naive estimate of allele frequencies by assuming the observed data represents the true underlying genotypes. We add 1 observation to each allele to ensure that there is positive probability of any allele. We sample proposed vectors of allele frequencies from a logistic normal distribution with mean π_j and an adaptive variance parameter that is also updated based on whether the previos proposal was accepted or not. We accept the proposal with probability

$$
\begin{aligned}
&= \min\{1, \frac{q(\pi_j|\pi_j^*)p(\pi_j^*|g_{\cdot,j}, \pi_j, \epsilon^-, \epsilon^+)}{q(\pi_j^*|\pi_j)p(\pi_j|g_{\cdot,j}, \pi_j, \epsilon^-, \epsilon^+)}\} \\
&= \min\{1, \frac{p(\pi_j^*|g_{\cdot,j}, \pi_j, \epsilon^-, \epsilon^+)}{p(\pi_j|g_{\cdot,j}, \pi_j, \epsilon^-, \epsilon^+)}\}
\end{aligned}
$$

due to symmetry of the proposal distribution.

False Positive and False Negative Rates

ϵ^+ and ϵ^- are initialized to user provided values. We sample proposed values of ϵ^+ and ϵ^- from logit normal distributions with means equal to the current values, and an adaptive

variance parameter that is updated based on whether the previous proposal was accepted or not. We accept the proposal with probability

$$
\begin{aligned}
&= \min\{1, \frac{q(\epsilon^+, \epsilon^-|\epsilon^{+^*}, \epsilon^{-^*})p(\epsilon^{+^*}, \epsilon^{-^*}|g, \pi, \mu)}{q(\epsilon^{+^*}, \epsilon^{-^*}|\epsilon^+, \epsilon^-)p(\epsilon^+, \epsilon^-|g, \pi, \mu)}\} \\
&= \min\{1, \frac{p(\epsilon^{+^*}, \epsilon^{-^*}|g, \pi, \mu)}{p(\epsilon^+, \epsilon^-|g, \pi, \mu)}\}
\end{aligned}
$$

due to symmetry of the proposal distribution.

Population Mean Complexity of Infection

λ is initialized to the mean of the initial values of μ. We sample from the posterior distribution of λ directly in a Gibbs Sampling step. The posterior distribution of λ can be expressed as

$$
\begin{aligned}
p(\lambda|\mu) &\propto p(\lambda)p(\mu|\lambda) \\
&= \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \prod_{i=1}^{n} \frac{\lambda^{\mu_i}}{(e^\lambda - 1)\mu_i!}
\end{aligned}
$$

which can be understood as sampling from

$$\lambda^* \sim \Gamma(\alpha + (\sum_{i=1}^{n} \mu_i) - n, \beta + n)$$

due to the conjugacy of the Zero-Truncated Poisson sampling distribution of $\mu|\lambda$ and the Gamma prior distribution of λ.

Likelihood Calculation

When calculating the likelihood of μ, π, ϵ^+, or ϵ^- above, we are required to calculate the probability of observing some relevant subset (or all) of the observed data conditional on the parameter, resulting in an expression of the form

$$
\begin{aligned}
p(g|\mu, \pi, \epsilon^+, \epsilon^-) &= \prod_{i=1}^{n} \prod_{j=1}^{l} p(g_{i,j}|\mu_i, \pi_j, \epsilon^+, \epsilon^-) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{l} \sum_{g^* \in G^*_{i,j}} p(g_{i,j}|g^*, \epsilon^-, \epsilon^+) p(g^*|\mu_i, \pi_j)
\end{aligned}
$$

# Possible Alleles	Complexity of Infection 1	2	3	4	5	6
2	2	3	4	5	6	7
4	4	10	20	35	56	84
8	8	36	120	330	792	1716
16	16	136	816	3876	15504	54264
32	32	528	5984	52360	376992	2324784
64	64	2080	45760	766480	10424128	119877472
128	128	8256	357760	11716640	309319296	6856577728

Table 2.2: Total Genotypes Possible

Each cell in the table indicates the total number of genotypes possible for a given complexity of infection and total number of alleles available at a single locus

We draw the reader's attention to the inner summation to marginalize over possible genotypes that could have resulted in the observed data. In the case of genotyping error allowing for both false positives and false negatives, the cardinality of the set $G^*_{i,j}$ is $\frac{(a_j+\mu_i-1)!}{\mu_i!(a_j-1)!}$. This quantity grows rapidly, quickly making computation intractable. Table 2.2 gives the total number of possible genotypes that must be considered under different combinations of total alleles and complexity of infection at a single genetic locus.

This poses a challenge to working with high diversity genetic loci, particularly in the context of high complexity infections when they are most useful. Under mild conditions [1], however, we may replace $p(g|\mu, \pi, \epsilon^+, \epsilon^-)$ with an estimator $\hat{p}(g|\mu, \pi, \epsilon^+, \epsilon^-)$ and still obtain exact inference for our model parameters [10]. This approach, known as the pseudo-marginal Metropolis-Hastings algorithm, allows us to formulate an estimator that remains computationally tractable while still maintaining the posterior distribution of the parameter as the stationary distribution of the MCMC.

In practice, we use importance sampling to estimate each $p(g_{i,j}|\mu_i, \pi_j, \epsilon^+, \epsilon^-)$ where the cardinality of $G^*_{i,j}$ exceeds some user defined threshold. The importance sampler is motivated by the following identity:

[1] The estimator of the target density in the Metropolis-Hastings acceptance ratio must be non-negative and unbiased

$$
\begin{aligned}
p(g_{i,j}|\mu_i, \pi_j, \epsilon^+, \epsilon^-) &= \sum_{g^* \in G^*_{i,j}} p(g_{i,j}|g^*, \epsilon^+, \epsilon^-) p(g^*|\mu_i, \pi_j) \\
&= \sum_{g^* \in G^*_{i,j}} p(g_{i,j}|g^*, \epsilon^+, \epsilon^-) \frac{p(g^*|\mu_i, \pi_j)}{q(g^*)} q(g^*) \\
&= \mathbb{E}[\frac{p(g_{i,j}|g^*, \epsilon^+, \epsilon^-) p(g^*|\mu_i, \pi_j)}{q(g^*)}] \\
\hat{p}(g_{i,j}|\mu_i, \pi_j, \epsilon^+, \epsilon^-) &= \frac{1}{S} \sum_{m=1}^{S} \frac{p(g_{i,j}|g^*_m, \epsilon^+, \epsilon^-) p(g^*_m|\mu_i, \pi_j)}{q(g^*_m)} \\
g^*_m &\sim q(\cdot) \text{ (Importance Sampling Distribution)} \\
\hat{p}(g|\mu, \pi, \epsilon^+, \epsilon^-) &= \prod_{i=1}^{n} \prod_{j=1}^{k} \hat{p}(g_{i,j}|\mu_i, \pi_j, \epsilon^+, \epsilon^-)
\end{aligned}
$$

Selection of an appropriate importance sampling density for sampling possible genotypes has a significant impact on the variance of the estimate, and subsequently on the convergence of the MCMC algorithm. We find that a multinomial distribution that incorporates the current estimate of the allele frequencies π_j, the observed data $g_{i,j}$, and the false negative rate ϵ^- performs well. We incorporate these parameters by taking the original estimate of the allele frequencies and reweighting them such that alleles that are not observed have a combined proportion equal to the false negative rate. For example, if $g_{i,j} = [1, 0, 0, 0]$, $\pi_j = [.1, .2, .3, .4]$ and $\epsilon^- = .1$, we would sample from a multinomial distribution with frequencies $[.9, .022, .033, .044]$.

Chapter 3

Simulation Study

We simulated 6 different data sets of 100 samples according to our model. We varied population mean COI over 3, 5, and 7, and 2 genetic panels consisting of 12 loci with 5 alleles each or 3 sets of 4 loci with 5, 15, and 25 alleles respectively. Allele frequencies were drawn from a symmetric Dirichlet distribution with $\alpha = 1$. ϵ^+ and ϵ^- were set to .03 and .1 respectively. The simulations were meant to represent the variety of environments encountered, ranging from moderate to high complexity of infection, and the range of technologies available, from relatively low diversity genetic loci to high diversity genetic loci, utilizing a genotyping method with moderate ability to detect alleles, and a false positive error rate on the high end of what would be tolerable. We compared the performance of our MCMC based approach to recovering COI and genetic locus allele frequencies across these different environments, as well as to the naive methods of estimating complexity of infection and allele frequencies.

Results

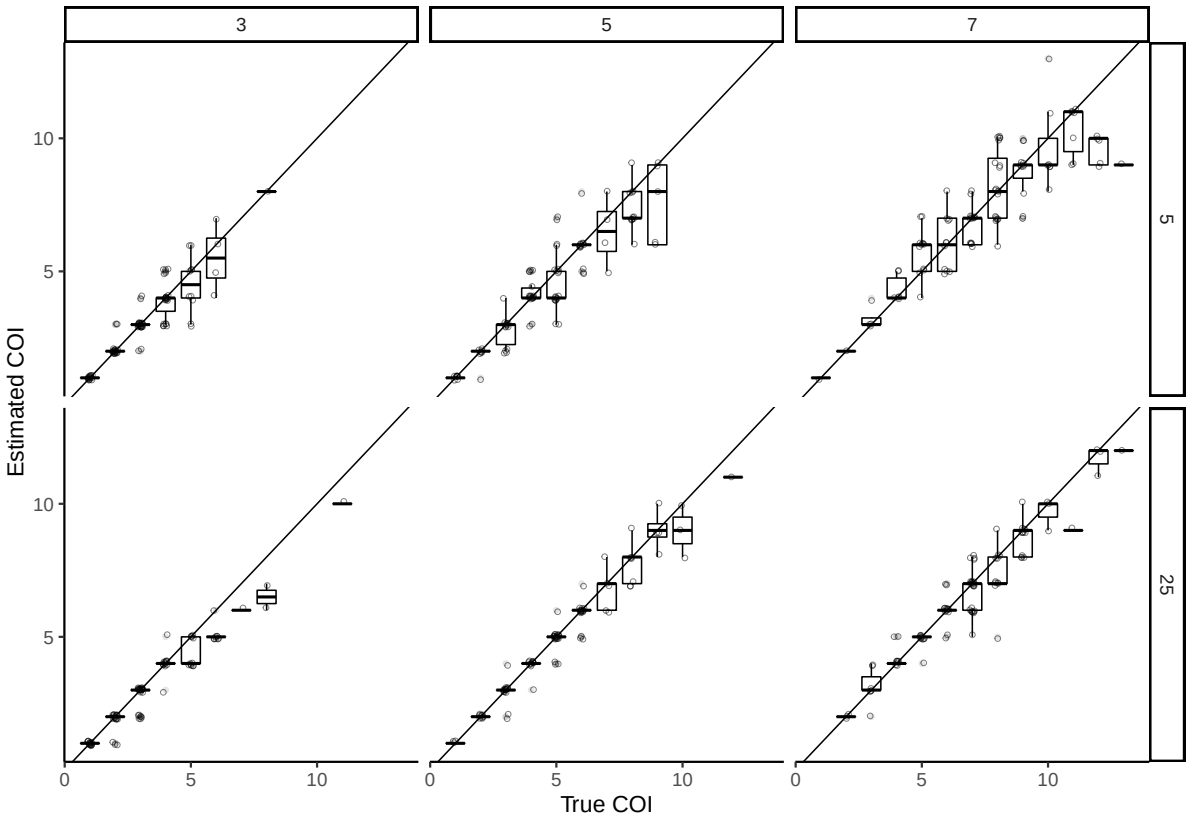

Figure 3.1: True vs Estimated COI

Each point represents the 50^{th} percentile posterior estimate of COI for one specimen, the thick line indicates the mean, and the edges of the boxes indicate the 25^{th} and 75^{th} percentiles. Mean complexity of infection is indicated at the top of the panel (3,5, or 7), and maximum number of alleles on the right side of the panel (5 or 25).

It is apparent that naive methods to estimation are heavily impacted by the diversity of the loci, resulting in extreme underestimation of COI in cases where genetic diversity is low but mean COI is high, and conversely very large overestimation of COI when genetic diversity is increased (Fig 3.2). This is further reflected in estimates of allele frequencies, where low frequency alleles are overestimated, and high frequency alleles are underestimated, leading to a general overestimate of genetic diversity (Fig 3.4). This underscores the biases that can be introduced using naive methods.

In contrast, we find that our method performs very well across the range of mean COIs (Fig 3.1). The addition of high diversity genetic loci, however, makes a substantial improvement on recovering COI in samples with very high COI. This is to be expected, and

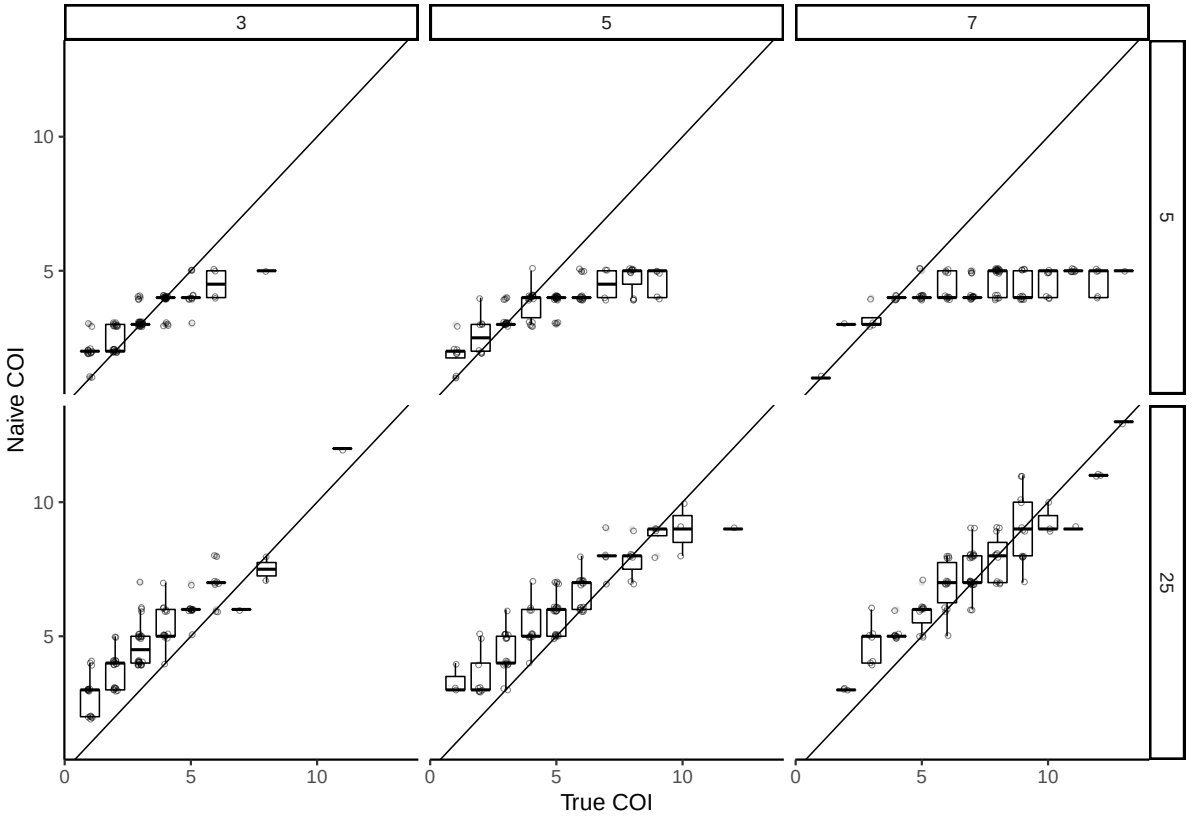

Figure 3.2: True vs Naively Estimated COI

Each point represents the naive estimate of COI for one specimen, the thick line indicates the mean, and the edges of the boxes indicate the 25^{th} and 75^{th} percentiles. Mean complexity of infection is indicated at the top of the panel (3,5, or 7), and maximum number of alleles on the right side of the panel (5 or 25).

our method scales well to these situations due to its importance sampling based likelihood estimation.

We also find that our method performs very well in estimating allele frequencies, even with highly diverse genetic loci (Fig 3.3). Performance remains largely the same between the low diversity and high diversity panel, which is to be expected as these data sets consisted of the same number of samples.

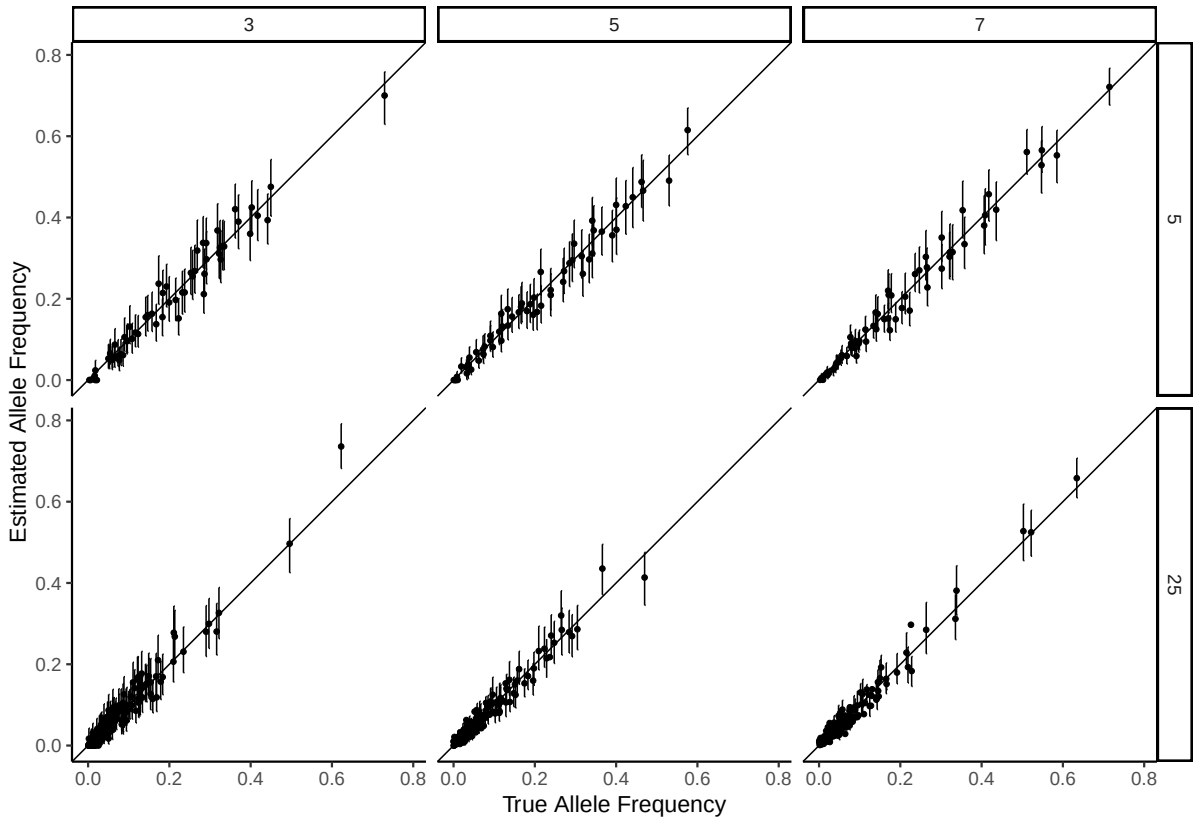

Figure 3.3: True vs Estimated Allele Frequency

Each point represents the 50^{th} percentile from the estimated posterior distribution for a single allele. The bars protruding indicate the 95% credible intervals for their estimates.

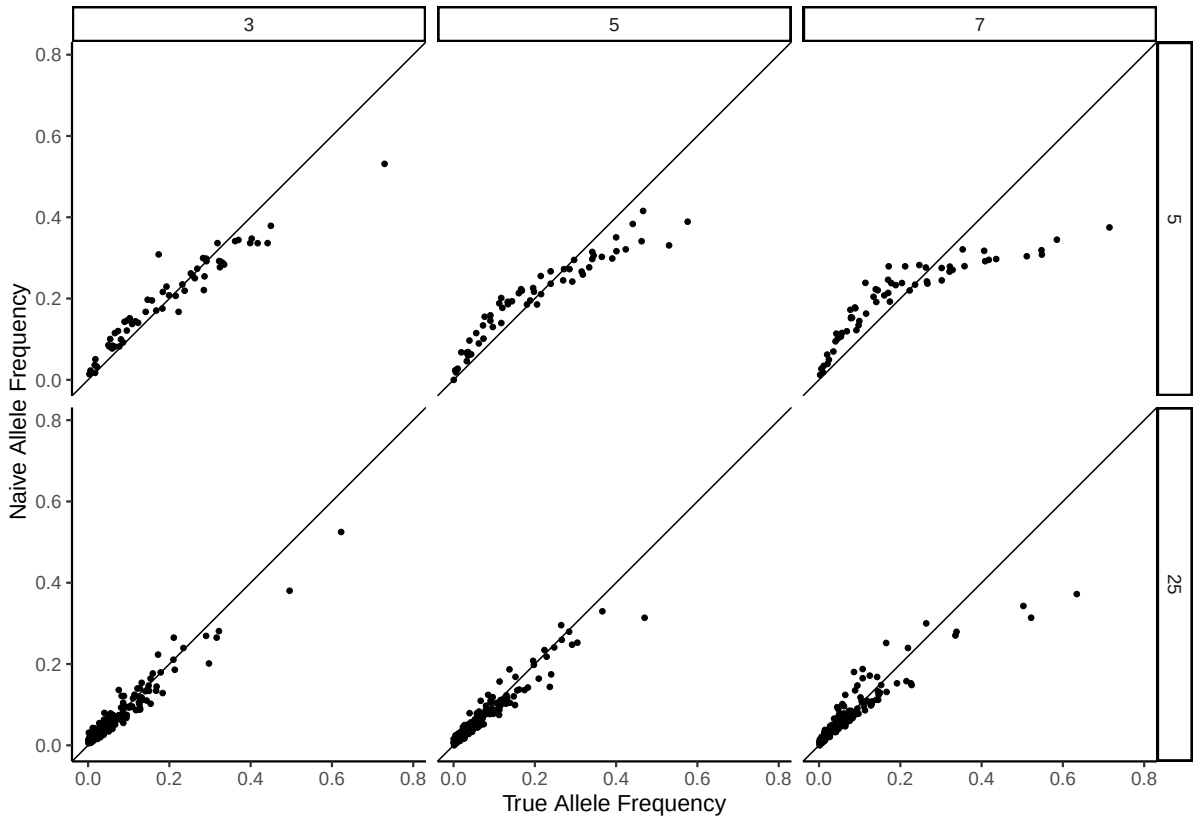

Figure 3.4: True vs Naively Estimated Allele Frequency

Each point represents the naive estimate of allele frequency for a single allele. Frequency was estimated by treating the indicator of an allele being present in the observed data as a single realization of that allele.

Chapter 4

Data Analysis

We next present an example of our method applied to a real world data set, demonstrating how interpretation is impacted. Our data come from passively and actively detected malaria cases in the Kingdom of Eswatini, formerly the Kingdom of Swaziland, from 2014 to 2016. Malaria transmission is very low in Eswatini, and is mostly sustained by imported infections. Samples were genotyped using 26 microsatellite markers by capillary electrophoresis. The microsatellite markers vary in diversity, ranging from 5 to 30 possible alleles. Table 4.1 lists the base sizes of alleles for each locus. We included 581 samples that were successfully genotyped at at least 23 of the 26 markers in the following analysis.

We fit our model by running multiple MCMC chains for $10,000$ iterations, discarding the first 5000 as burn-in, and assessed convergence by comparing across chains, ensuring that they had similar posterior distribution estimates. We used weakly informative $Beta(1, 99)$ and $Beta(5, 95)$ priors over ϵ^+ and ϵ^- to constrain to low values and set an importance sampling threshold of 1000. We estimated mean complexity of infection, sample complexity of infection, genetic locus allele frequencies and functions thereof by the 50^{th} percentile estimate of the respective marginal posterior distributions. All estimates are given with 95% credible intervals. We contrast these estimates with other widely used estimation techniques. These include:

- Estimation of μ by the maximum number of observed alleles across genetic loci for each sample, which we refer to as the **Naive COI** method.
- Estimation of μ by the second maximum number of observed alleles across genetic loci for each sample, attempting to acknowledge the inherent bias of this method, which we refer to as the **Naive Offset COI** method.
- Estimation of π by treating each observation in g as a single allele realization, which we refer to as the **Naive Allele Frequency** method.

Locus	Alleles	# Alleles
Ara2	120, 124, 130, 133, 136, 139, 143, 146, 149, 152, 155, 158, 161	13
AS1	167, 170, 173, 176, 179, 182, 185	7
AS2	181, 184, 187, 189, 192, 195, 198, 202, 204, 208, 211	11
AS3	158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 191	11
AS7	156, 162, 164, 168, 171, 174, 177, 179, 182, 185, 188, 194, 196	13
AS8	191, 193, 198, 201, 204, 219, 223, 226	8
AS11	143, 146, 150, 153, 156, 159, 162, 165, 168, 173, 174, 176, 180, 183	14
AS12	156, 162, 165, 168, 171	5
AS14	187, 193, 197, 200, 203, 206, 209, 212, 215, 218, 221, 224, 227, 230, 233	15
AS15	112, 116, 119, 122, 125, 128, 131, 134, 137, 140, 143, 146, 149, 153, 156	15
AS19	133, 136, 139, 143, 146, 149, 153, 157, 160, 162, 167, 173, 176, 179, 182, 185, 188	17
AS21	146, 154, 157, 160, 163, 176, 179, 182	8
AS25	67, 73, 77, 81, 85, 88, 91, 94, 98, 101, 104, 107, 110, 113, 117, 120, 123, 126, 129, 132, 139, 142, 148, 152	24
AS31	156, 159, 162, 165, 168, 171, 173, 176, 180, 184, 187, 191, 194, 197, 200, 203, 206, 209, 212, 215, 218, 221, 224, 227	24
AS32	205, 211, 223, 227, 230, 233, 236, 239, 242, 245, 248, 251, 258, 261, 263, 267, 273, 276, 288	19
AS34	162, 178, 181, 184, 187, 190, 193	7
B7M19	138, 147, 155, 164, 172, 181	6
PFG377	94, 97, 100, 103, 107, 110, 113	7
PfPK2	151, 155, 158, 167, 170, 173, 176, 179, 182, 185, 188, 191, 194, 197, 200, 203, 206, 209, 212	19
PolyA	99, 109, 112, 115, 118, 121, 124, 126, 130, 136, 140, 143, 146, 149, 152, 155, 158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 188, 191, 197, 200	30
TA1	130, 134, 137, 140, 143, 147, 153, 156, 159, 162, 165, 168, 171, 174, 177, 181, 184, 187, 190, 193, 196, 199, 202, 205, 208, 211, 227	27
TA40	228, 232, 235, 238, 241, 244, 247, 250, 253, 256, 259, 262, 265, 268, 271, 274, 277, 280, 283, 286, 289, 292, 295, 298	24
TA60	199, 202, 205, 209, 212, 215, 218, 221, 224, 227, 247, 250	12
TA81	110, 113, 115, 118, 121, 125, 128, 131, 134, 137, 140, 143, 146, 155, 162, 165	16
TA87	82, 85, 88, 91, 94, 97, 100, 103, 106, 110, 113, 116, 119, 122, 125, 128, 136	17
TA109	140, 143, 146, 149, 152, 155, 158, 162, 165, 168, 171, 174, 177, 180, 183, 186, 189, 193, 195, 199, 202	21

Table 4.1: Microsatellite Markers

COI Estimation

By our method, we estimated the mean COI to be 1.61 $(1.54 - 1.68)$. By the Naive COI method and Naive Offset COI method, we estimated the mean COI to be 2.66 and 2.12 respectively. Distributions of estimated COI are displayed in figure 4.1. By our method, 42.8% $(42.3\% - 43.5\%)$ of samples were considered polyclonal, while the Naive and Naive Offset methods estimated 74.7% and 60.9% to be polyclonal respectively. This is a substantial difference in estimates and would impact understanding of malaria transmission dynamics in this setting. These results also reflect what we observed in simulations, that COI is consistently overestimated by the naive method when multiple diverse loci are used.

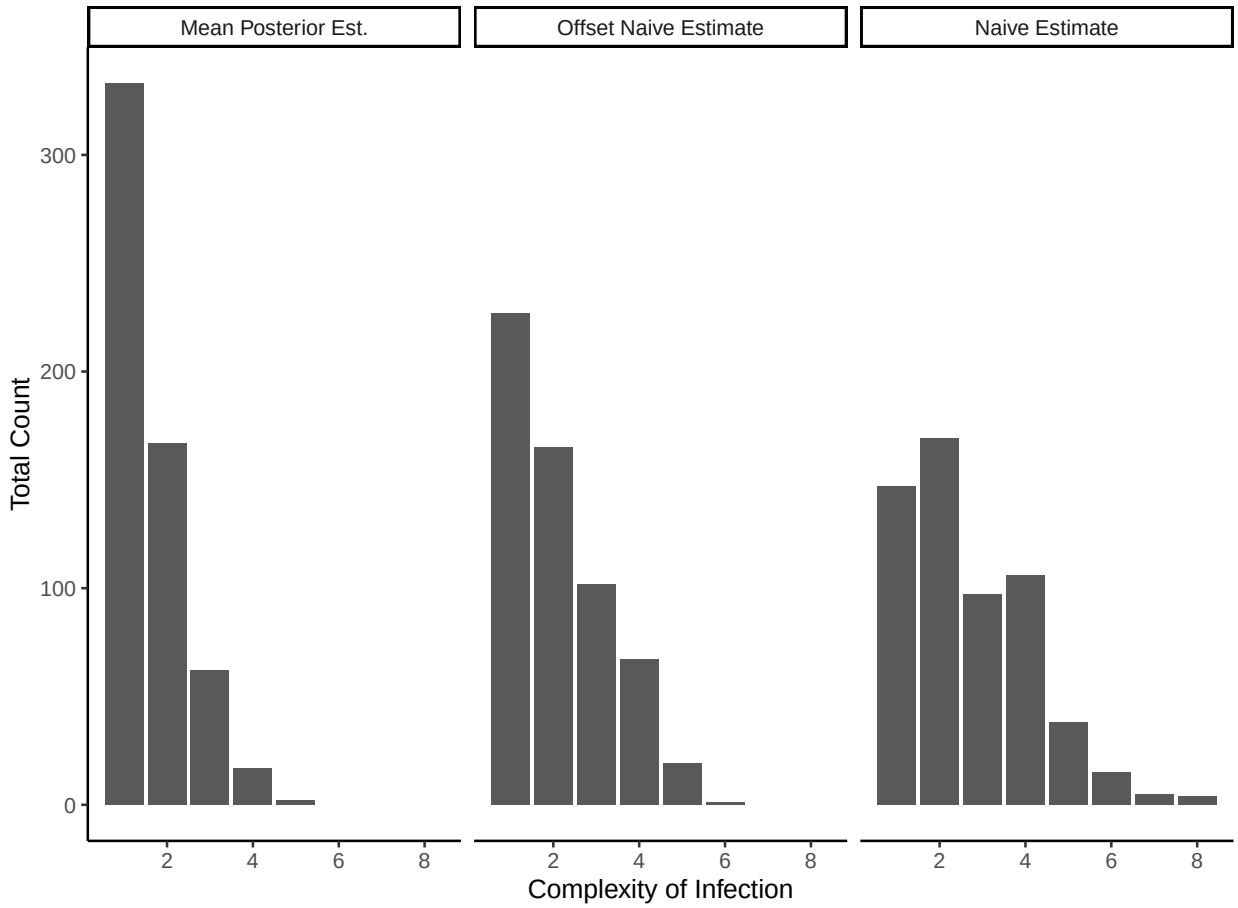

Figure 4.1: Distributions of Estimated COI by Multiple Methods

Each panel displays the distribution of estimated COIs across different methods

Allele Frequencies and Genetic Diversity

Population level genetic diversity was explored by estimating expected heterozygosity (H_E). H_E is defined as the probability of randomly drawing two different alleles from the available allele pool. H_E ranges from 0 indicating no diversity to 1, indicating every allele is different. Heterozygosity for a given locus is maximized when every allele is equally likely to be observed. We may estimate H_E for a given locus by the following equation [2]:

$$H_E = \frac{n}{n-1}[1 - \sum_{k=1}^{a_j} \pi_{j,k}^2]$$

By our method, we estimated mean H_E across loci as .707 (.703 – .711). By the Naive Allele Frequencies method, we estimated mean H_E to be .743. Estimates of heterozygosity at each locus by both methods are shown in figure 4.2. This illustrates a trend of the

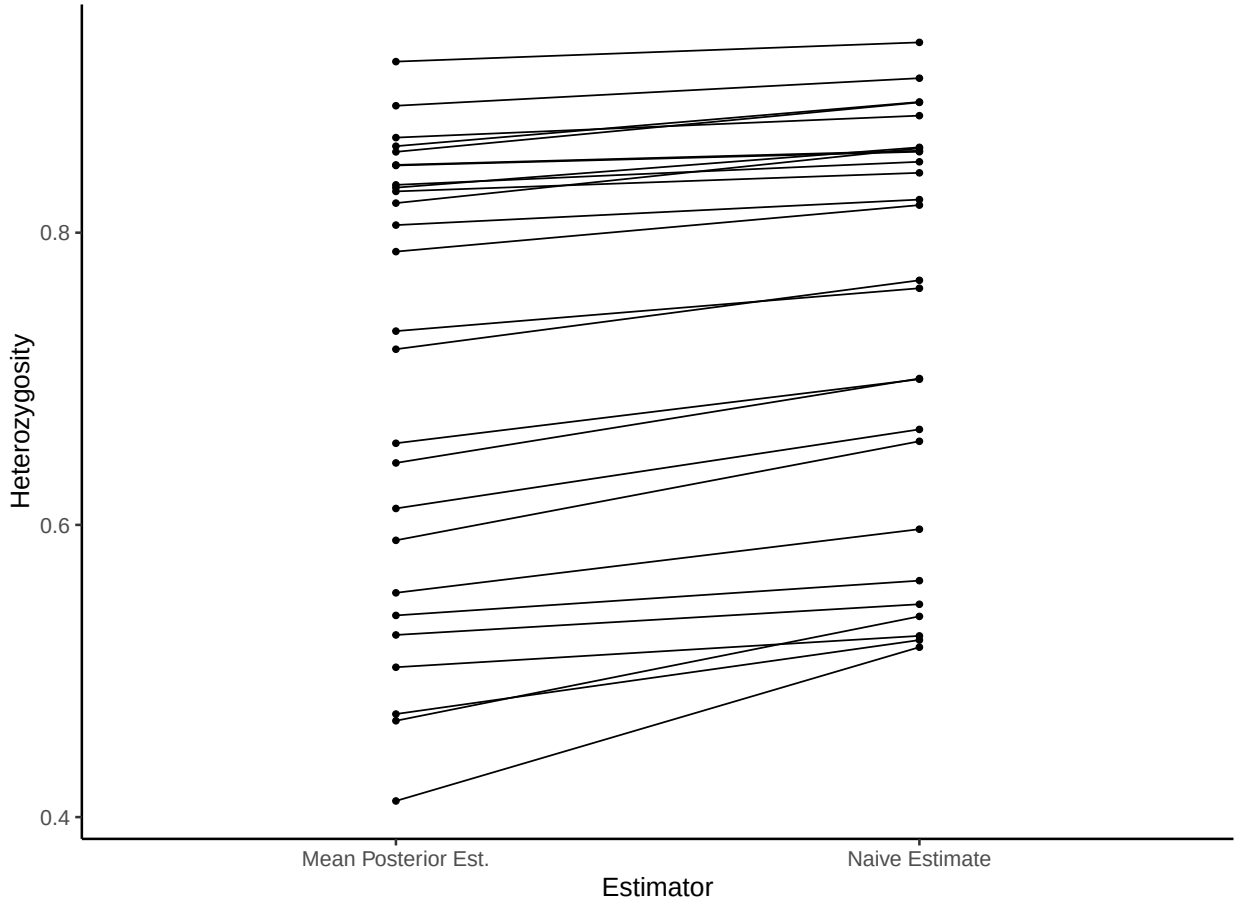

Figure 4.2: Estimates of Heterozygosity For Each Locus

Each dot represents the estimate of heterozygosity for a single locus.

naive method estimating higher genetic diversity across loci, concordant with our simulations before that shows that naive estimation underestimates common allele frequencies and overestimates rare allele frequencies.

Chapter 5

Conclusion

With this thesis, we have developed and demonstrated a new model for estimating complexity of infection and genetic locus allele frequencies from multi-allelic data. This is an important step forward as genotyping methods shift towards panels of more diverse genetic markers, allowing for much higher resolution genetics based analysis. Further, our approach to estimating likelihoods for MCMC proposal updates has greatly expanded the model's applicability by addressing a computational burden that was otherwise insurmountable. We have demonstrated its performance using simulations, showing that it is able to recover accurate estimates of COI and allele frequencies. We have also demonstrated on real data how using this model can lead to substantially different estimates of epidemiologically and ecologically relevant parameters such as proportion of polyclonal infections and mean heterozygosity compared to the widely used biased methods.

Software Availability

In order to facilitate widespread use of this method in the malaria genetics community, we have made an **R** package available. The software is written using **Rcpp**, making the code fast and efficient. Details of the software and how to use and install it are available at https://www.github.com/m-murphy/moiR

Future Directions

We hope to extend this model further and address and weaken some of the assumptions that are currently made. Possible extensions that are of interest include: incorporating spatial information to inform allele frequencies, rather than assuming a fixed, well mixed population; allowing for other less rigid distributions over complexity of infection; extending the error model to allow for differences across genetic markers, as well as be informed by other factors such as sample quality; and finally incorporate estimates of relative abundances of DNA at

observed alleles to further improve accuracy and potentially estimate phasing of genotypes into their constituent distinct parasite strains.

Bibliography

[1] Joshua Adjah et al. "Seasonal variations in Plasmodium falciparum genetic diversity and multiplicity of infection in asymptomatic children living in southern Ghana." In: *BMC Infectious Diseases* 18.1 (Aug. 2018), p. 432. URL: http://dx.doi.org/10.1186/s12879-018-3350-z.

[2] T J Anderson et al. "Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum." In: *Molecular Biology and Evolution* 17.10 (Oct. 2000), pp. 1467–1482. DOI: 10.1093/oxfordjournals.molbev.a026247. URL: http://dx.doi.org/10.1093/oxfordjournals.molbev.a026247 (visited on 05/10/2019).

[3] Sarah Auburn and Alyssa E Barry. "Dissecting malaria biology and epidemiology using population genetics and genomics." In: *International Journal for Parasitology* 47.2-3 (2017), pp. 77–85. DOI: 10.1016/j.ijpara.2016.08.006. URL: http://dx.doi.org/10.1016/j.ijpara.2016.08.006.

[4] Oliver Balmer and Marcel Tanner. "Prevalence and implications of multiple-strain infections." In: *The Lancet Infectious Diseases* 11.11 (Nov. 2011), pp. 868–878. DOI: 10.1016/S1473-3099(11)70241-9. URL: http://dx.doi.org/10.1016/S1473-3099(11)70241-9.

[5] Hsiao-Han Chang et al. "THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites." In: *PLoS Computational Biology* 13.1 (Jan. 2017), e1005348. DOI: 10.1371/journal.pcbi.1005348. URL: http://dx.doi.org/10.1371/journal.pcbi.1005348.

[6] Jiang-Tao Chen et al. "Genetic diversity and allele frequencies of Plasmodium falciparum msp1 and msp2 in parasite isolates from Bioko Island, Equatorial Guinea." In: *Malaria Journal* 17.1 (Dec. 2018), p. 458. DOI: 10.1186/s12936-018-2611-z. URL: http://dx.doi.org/10.1186/s12936-018-2611-z.

[7] Bo Huang et al. "Temporal changes in genetic diversity of msp-1, msp-2, and msp-3 in Plasmodium falciparum isolates from Grande Comore Island after introduction of ACT." In: *Malaria Journal* 17.1 (Feb. 2018), p. 83. URL: http://dx.doi.org/10.1186/s12936-018-2227-3.

[8] Bruno Moonen et al. "Operational strategies to achieve and maintain malaria elimination." In: *The Lancet* 376.9752 (Nov. 2010), pp. 1592–1603. DOI: 10.1016/S0140-6736(10)61269-X. URL: http://dx.doi.org/10.1016/S0140-6736(10)61269-X.

[9] Sílvia Portugal, Hal Drakesmith, and Maria M Mota. "Superinfection in malaria: Plasmodium shows its iron will." In: *EMBO Reports* 12.12 (Dec. 2011), pp. 1233–1242. DOI: 10.1038/embor.2011.213. URL: http://dx.doi.org/10.1038/embor.2011.213.

[10] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. 2004. Corr. 2nd. Springer-Verlag New York, LLC, Nov. 2010, p. 679. ISBN: 978-1-4419-1939-7.

[11] Michelle E Roh et al. "High genetic diversity ofPlasmodium falciparum in the low transmission setting of the Kingdom of Eswatini: Supplementary Appendix". In: *BioRxiv* (Jan. 2019). DOI: 10.1101/522896. URL: http://biorxiv.org/lookup/doi/10.1101/522896.

[12] Than Naing Soe et al. "Genetic diversity of Plasmodium falciparum populations in southeast and western Myanmar." In: *Parasites & vectors* 10.1 (July 2017), p. 322. DOI: 10.1186/s13071-017-2254-x. URL: http://dx.doi.org/10.1186/s13071-017-2254-x.

[13] *WHO World malaria report 2018*. WEBSITE. URL: https://www.who.int/malaria/publications/world-malaria-report-2018/en/ (visited on 02/12/2019).